Datawarehouse delle serie storiche R&S in Italia

Nicola Ferrante
CILEA, Roma

Abstract
Questo articolo illustra e documenta la progettazione di un sistema che permette analisi di tipo OLAP (On-Line Analytical Processing). L'obiettivo è stato quello di creare strumenti che permettessero di ottenere in tempi ridotti informazioni riassuntive, che potessero essere di aiuto agli analisti nell'ambito del progetto FIRB tra CILEA, ISTAT, CNR e ConfIndustria. Per la realizzazione di questo datawarehouse saranno proposte due soluzioni alternative, diverse dal punto di vista tecnologico e metodologico: un approccio più conservatore (ROLAP, Relational OLAP) messo a confronto con uno più moderno (MOLAP, Multidimensional OLAP).

This paper illustrates and documents the planning of a system that allows OLAP (On-Line Analytical Processing) analysis. The objective has been to create instruments that allowed to obtain quickly summarised informations that could be of aid to the analysts in the within of FIRB project between CILEA, ISTAT, CNR and Confindustria. In order to realise this datawarehouse two solutions will be proposed, various from the technological and methodological point of view: a more conservative approach (ROLAP, Relational OLAP) compared with a more modern one (MOLAP, Multidimensional OLAP).

Keywords: Datawarehouse, OLAP, ROLAP, MOLAP, Oracle, Business intelligence.

Introduzione

Con le rilevazioni annuali sulla ricerca e sviluppo (R&S) l'STAT raccoglie informazioni statistiche presso imprese, amministrazioni pubbliche e istituzioni private no-profit circa l'attività di R&S svolta al loro interno. I principali indicatori presentati in queste tavole riguardano la spesa sostenuta per attività di R&S (R&S *intra-muros*) e il personale impegnato in attività di R&S. La rilevazione sulla Ricerca e lo Sviluppo sperimentale (R&S) in Italia viene condotta dall'ISTAT a partire dal 1963; attualmente vengono condotte tre indagini separate: una riguardante le imprese (RS1), una relativa alle amministrazioni pubbliche (RS2) e una riguardante le istituzioni private no-profit (RS3).

Il progetto offre la possibilità di studiare l'impatto degli investimenti (pubblici e privati) sui risultati produttivi e competitivi delle imprese, usando tassonomie di comportamenti innovativi e di uso di supporti pubblici.

Nell'ambito di un progetto FIRB sulla valutazione dell'impatto degli investimenti in ricerca e innovazione nelle imprese sul sistema produttivo del paese [1], ci si è posti l'obiettivo di creare strumenti per ottenere in tempi ridotti informazioni riassuntive che possano essere di aiuto agli analisti, nello specifico, realizzare un datawarehouse.

Il datawarehouse

Secondo W.H.Inmon, un datawarehouse è "un insieme di dati subject oriented, integrato, time variant, non volatile costruito per supportare il processo decisionale" [2]. Per la realizzazione di questo database statistico saranno proposte due soluzioni alternative, diverse dal punto di vista tecnologico e metodologico. Verrà proposta una prima soluzione usando la tecnologia relazionale (*Relational-OLAP* o *ROLAP*) in cui i dati vengono memorizzati tramite tabelle e le operazioni di analisi tradotte in opportune istruzioni SQL eseguite dal dbms (nel nostro caso Oracle 10g) utilizzando strutture relazionali per l'accesso ai dati.

Una seconda soluzione metodologicamente diversa dalla prima (*Multidimensional-OLAP* o *MOLAP*), realizzata attraverso gli strumenti che offre la tecnologia *Oracle 10g*. In questo caso i dati sono memorizzati direttamente in forma

multidimensionale tramite strutture dati proprietarie. Si illustreranno i passi per la creazione del modello multidimensionale con questa tecnologia (*Oracle Workspace Manager 10g*), quindi la realizzazione delle tabelle dei fatti, delle tabelle dimensioni e dei data mart (chiamati "cubi" in quest'ambito). Successivamente si proporrà una soluzione per la creazione dinamica dei report grafico-tabellari con lo strumento CASE *Oracle Business Intelligence Discoverer with OLAP*.

Secondo la definizione di datawarehouse, questo strumento si caratterizza soprattutto come collezione di dati a supporto del processo decisionale del management. Il datawarehouse raggruppa i dati decisionali per aree o temi di interesse e li organizza rispetto all'utilizzazione finale; si differenzia in questo dai tradizionali database la cui progettazione è guidata dai requisiti delle applicazioni che garantiscono i processi gestionali. In ambito datawarehouse le informazioni assumono un valore aziendale piuttosto che dipartimentale. Si ha un orizzonte temporale più ampio, garantendo il mantenimento di informazioni storiche, in modo da poter favorire le attività di analisi comparative su diversi periodi temporali. Le informazioni sono consolidate, consistenti nel tempo e non modificabili dall'utente che le accede esclusivamente in lettura. I sistemi di datawarehousing consentono di acquisire e integrare informazioni provenienti da sorgenti eterogenee e di interrogare efficientemente basi di dati di grandi dimensioni.

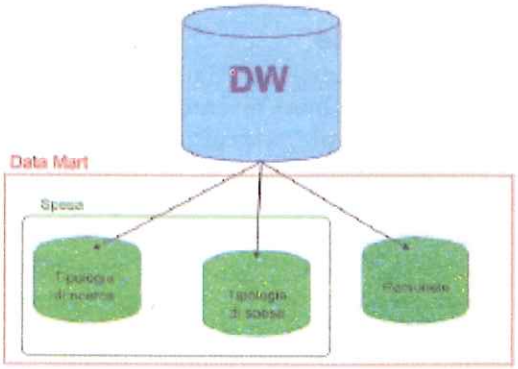

Fig. 1 – Struttura del datawarehouse

Dal punto di vista funzionale, il processo di datawarehousing consiste in tre fasi: estrazione dei dati da sorgenti operazionali distribuite e loro integrazione; organizzazione dei dati nel datawarehouse; accesso efficiente e flessibile ai dati integrati. La terza fase richiede capacità di navigazione degli aggregati, ottimizzazione di interrogazioni complesse, tecniche di indicizzazione avanzate e interfacce visuali amichevoli per l'OLAP.

Nella letteratura è universalmente riconosciuto che un datawarehouse utilizza un modello multidimensionale dei dati (*Dimensional Fact Model* o *DFM* [3] [4]). La rappresentazione costruita tramite il DFM è detta schema dimensionale e consiste in un insieme di schemi di *fatto*[1] i cui elementi costituenti sono *misure*[2], *dimensioni*[3] e *gerarchie*[4].

Datawarehouse delle Serie storiche R&S in Italia

Nel nostro ambito possiamo inquadrare più *data mart*, cioè archivi differenziati per obiettivo, settore oppure oggetto di analisi. Le tabelle dimensione sono generalmente denormalizzate[5]. I tre data mart si differenziano per ambito di analisi (fig. 1):

- *tipologia di spesa*: misurata in base a spese correnti, spese del personale e spese in conto capitale;
- *tipologia di ricerca*: contraddistinta in ricerca di base, ricerca applicata e sviluppo;
- *composizione del personale*: si propone di supportare l'analisi sulla tipologia del personale misurato numericamente (head-count) o in tempo pieno equivalente (*full time equivalent*).

[1] Il concetto sul quale ha senso svolgere un processo di analisi (la spesa, i finanziamenti, etc.); è ciò su cui si incentra il processo decisionale, esso modellizza un evento che accade nella realtà modellizzata.

[2] Un attributo (tipicamente numerico) a valori continui che descrive un fatto sotto un determinato punto di vista. Solitamente il fatto possiede più misure.

[3] Una particolare prospettiva lungo la quale l'analisi di un fatto può essere effettuata. Le dimensioni sono attributi discreti, che determinano la granularità minima adottata per rappresentare il fatto.

[4] Sono costituite da attributi a valori discreti legati da relazioni molti a uno e determinano in che modo i fatti possono essere aggregati e selezionati nell'ambito del processo decisionale. Per esempio, nella gerarchia *giorno? mese? anno* la radice della gerarchia, (giorno) rappresenta la granularità più fine (il massimo livello di dettaglio); mentre le altre due corrispondono a un livello sempre più inferiore di dettaglio.

[5] Contengono le dipendenze funzionali delle gerarchie, tutto ciò sveltisce le interrogazioni.

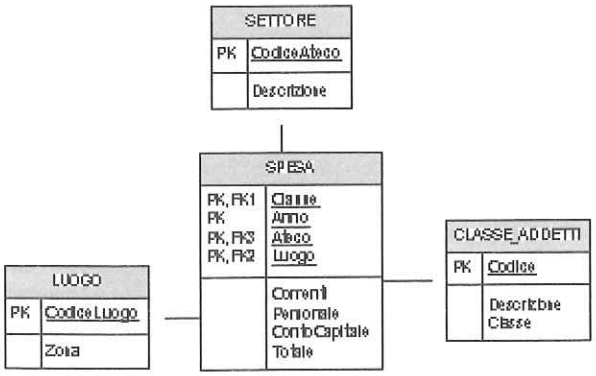

Fig. 2 – Schema a stella del data mart "tipologia di spesa"

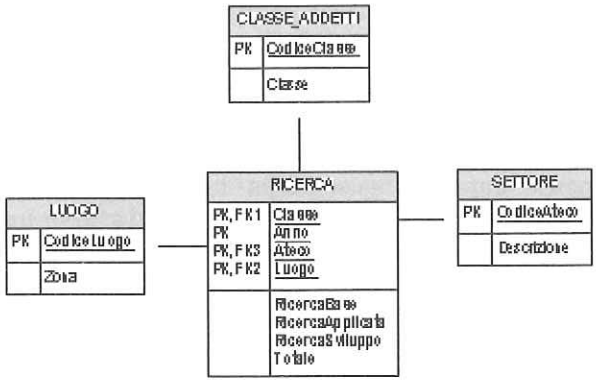

Fig. 3 – Schema a stella del data mart "tipologia di ricerca"

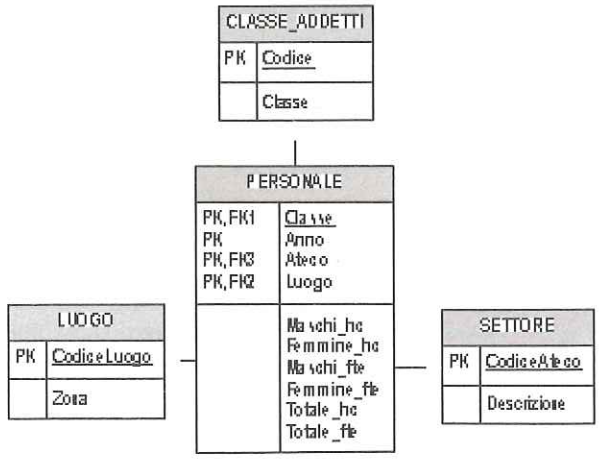

Fig. 4 – Schema a stella del data mart "composizione del personale"

Il processo di acquisizione [5] è normalmente la componente più costosa e complessa, sia nella fase di sviluppo sia in quella di manutenzione del datawarehouse [6]. Per acquisizione si intende il processo nell'ambito del quale sono svolte le seguenti attività [7]:

- *estrazione dei dati*: l'acquisizione dai sistemi operazionali e da fonti esterne dei dati utili al processo decisionale;
- *pulizia dei dati*: il processo di *cleaning*, che mira alla omogeneizzazione dei dati, può evidenziare carenze e/o incompletezze che inducono un processo di correzione a livello dei dati sorgente;
- *trasformazione:* attività di riorganizzazione e aggregazione di dati. I dati vengono inoltre arricchiti attraverso elaborazioni fatte rispetto alle diverse dimensioni di aggregazione, in particolare per la caratterizzazione della dimensione temporale;
- *caricamento*: i dati estratti e trasformati vengono caricati sui data mart per consentire l'analisi da parte degli addetti.

Un approccio Relational-OLAP

Il compito della creazione dell'ambiente R-OLAP è affidato al software *DWLoader* (un componente ETL progettato e realizzato appositamente con le seguenti caratteristiche:

1. creazione dell'ambiente ROLAP per il datawarehouse R&S su un dato schema e istanza di database (che è possibile specificare dall'interfaccia grafica);
2. inizializzazione delle tabelle di utilità e delle dimensioni che necessitano di dati da sorgenti pubbliche;
3. estrazione, pulizia, trasformazione e caricamento dei dati da file excel e importazioni sulle tabelle *wrapper* del datawarehouse.

La principale fonte di dati è data dai rapporti che l'ISTAT invia al CILEA su file excel. Prima di tutto il componente crea l'ambiente nel ROLAP, quindi procede con la creazione delle tabelle con i rispettivi vincoli e indici. Crea le dimensioni e le tabelle di utility (*wrappers*) grazie ai file di DUMP (delle sorgenti pubbliche e interne) che ha a disposizione; successivamente effettua il *parsing* dei file excel e l'elaborazione dei dati in essi contenuti. Infine trasferisce i dati su una serie di tabelle wrapper di un database relazionale (fig. 5).

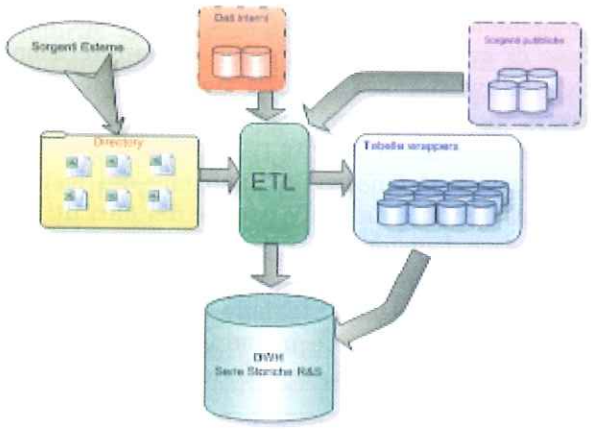

Fig. 5 – Diagramma di flusso dinamico dei dati

Le tabelle *wrapper* sono state pensate come struttura dati stabile, su cui procedere con l'elaborazione, la pulizia e la trasformazione dei dati. In questo modo è possibile elaborare i dati a livello di database attraverso *query* di inserimento, modifica o *store*. A questo punto vengono eseguite delle procedure che effettuano un'ulteriore pulizia dei dati, per rendere efficiente la fase di trasformazione.

Nella fase di trasformazione vengono generate le chiavi surrogate, che poi saranno inserite nelle *fact table* e nelle tabelle dimensione nella fase di caricamento. Nella fase di caricamento il software individua le *tuple* che dovranno essere aggiornate (*update*) o inserite (*insert*).

Per quanto riguarda la terza fase, ovvero l'analisi vera e propria sui dati, è stata progettata e realizzata un'applicazione web. Tale applicazione deve poter eseguire analisi dimensionale a partire da un data mart e dalle dimensioni. L'utente, dopo essersi opportunamente autenticato, deve poter selezionare il data mart da analizzare, nonché i valori di interesse per le dimensioni (fig. 7).

L'applicazione è formata da 3 livelli (fig. 6), ciascuno dei quali comprende una serie di componenti i quali hanno ruoli funzionali ben diversi. Particolare attenzione va data al componente datawarehouse della logica applicativa, nel senso che è stato sviluppato attraverso un massiccio uso di *reflection object*, i quali consentono di ottenere le informazioni relative ai tipi contenuti in un *assembly* a run-time.

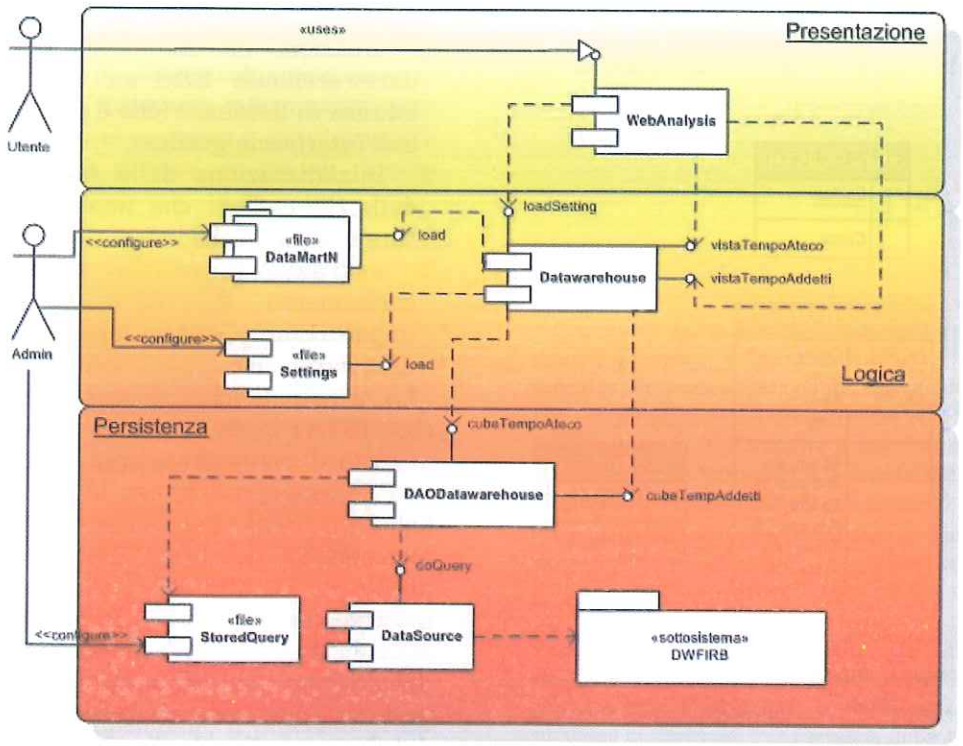

Fig. 6 – I livelli dell'applicazione web che gestisce l'analisi sui dati

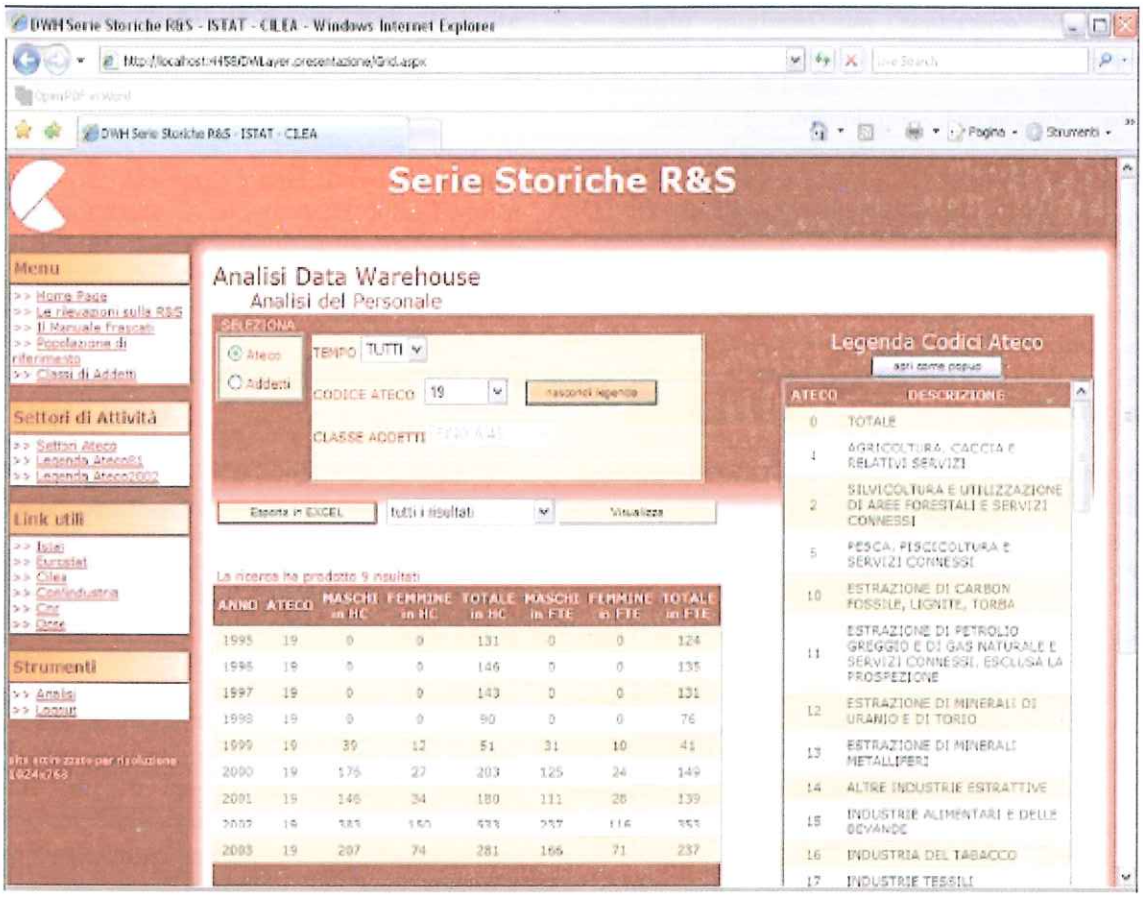

Fig. 7 – Snapshot di una pagina di analisi su web

Perciò, per esempio, qualora si voglia aggiungere (eliminare o modificare) una dimensione in uno dei data mart o una misura in una tabelle fatto (o anche una dimensione), non si dovrà modificare e ricompilare il codice con le nuove proprietà, ma basterà inserire le nuove configurazioni in un opportuno file e riavviare l'applicazione [8].

Tutto ciò consente a un amministratore (potenzialmente diverso dallo sviluppatore) di modificare gli schemi logici a seconda dell'utilità, senza coinvolgere altri soggetti e senza ricompilare il codice sorgente.

Un approccio Multidimensional-OLAP

La stragrande maggioranza dei sistemi MOLAP, nel nostro caso la tecnologia Oracle 10g [9], possiedono una fornitissima libreria di funzioni finanziarie e statistiche o ancora, incorporano una sofisticata, ma trasparente, gestione del reporting e dei grafici. In questo paragrafo si propone per il datawarehouse R&S una una soluzione con metodologia MOLAP. Verranno perciò discusse le potenzialità e l'utilizzo di tali strumenti, sia di creazione che di reporting, utilizzando principalmente due strumenti: Oracle Workspace Manager 10g e ORACLE Business Intelligence Discoverer with OLAP.

ORACLE Workspace Manager 10g offre la possibilità di definire uno strato logico multidimensionale. Questo strumento permette di definire i data mart in modo appropriato con le relative dimensioni (con tutti i rispettivi livelli e gerarchie), le misure (consente non solo di mappare campi del database nella struttura logica OLAP, ma anche di aggiungere altre misure calcolate dinamicamente). Le dimensioni possono avere uno o più livelli e sono organizzate in una o più gerarchie. Le gerarchie definiscono come i dati possano essere raggruppati a seconda del livello di astrazione richiesto da chi vuole fare l'analisi.

L'opzione di OLAP di Oracle 10g fornisce un modello dimensionale logico e la capacità di memorizzare i dati sia nei *datatype* relazionali che multidimensionali e fornisce molti vantaggi (anche di prestazioni) in confronto a una implementazione relazionale pura.

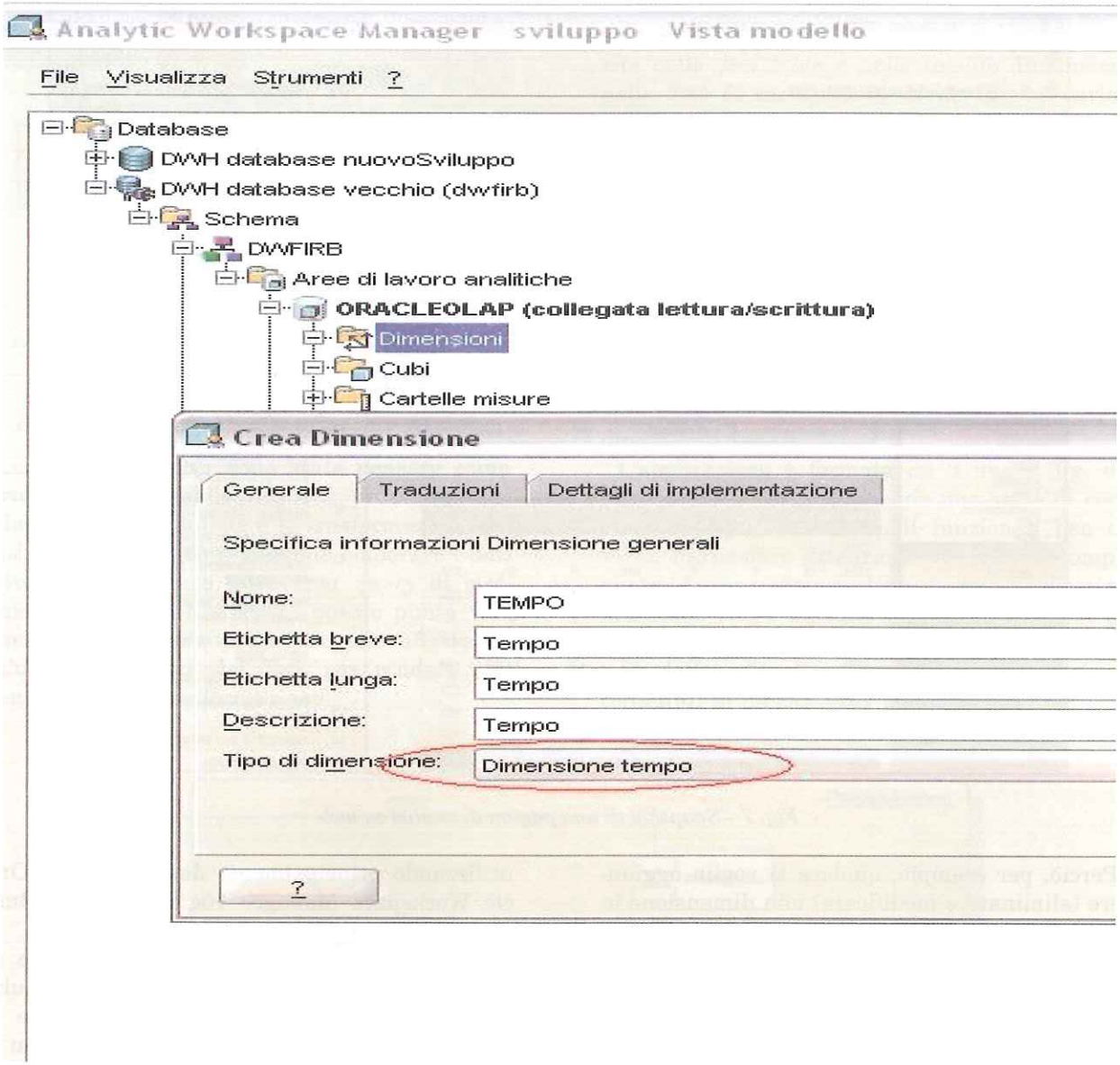

Fig. 8 – Snapshot della creazione di una dimensione in Oracle Workspace Manager 10g

I *datatype* dimensionali memorizzano i dati in strutture che offrono benefici di performance in un ambiente di analisi creato ad-hoc, il che rende il processo di applicare condizioni alle query estremamente efficiente. Il motore multidimensionale di *OLAP* option for ORACLE 10g contiene le ottimizzazioni per la gestione degli insiemi di grandi moli di dati, nonché sparsi; inoltre offre automaticamente supporto alla navigazione fra i dati di basso-livello e le successive aggregazioni di gerarchie.

Attraverso ORACLE Workspace Manager 10g si possono creare dimensioni, data mart, misure e piani di calcolo.

Per esempio, nel creare una dimensione innanzitutto si sceglie un'etichetta, dopodiché bisogna specificare il tipo di dimensione (fig. 8), che può essere o dimensione utente o dimensione tempo (ORACLE fornisce automaticamente il supporto per aggregare i vari livelli di granularità di questo tipo di dimensione). Inoltre, si può scegliere se utilizzare le chiavi già presenti nell'origine dati o generare chiavi univoche alternative (possibile solo nel caso in cui siano specificati più livelli nella gerarchia di quella dimensione).

L'ultimo passo da eseguire per implementare la dimensione è quello di eseguire il mapping logico (fig. 9) tra le tabelle del database e la

struttura logica OLAP. Il mapping che l'utente fa ad alto livello si traduce, a livello di DBMS, in una tabella di *metadati*[6], che descrivono le relazioni tra le entità logiche e permettono, dopo un eventuale aggiornamento della base di dati OLTP, di caricare i nuovi dati in modo da non compromettere la consistenza della base di dati OLAP. Al crescere delle informazioni memorizzate nel datawarehouse e all'aumentare delle modalità con cui queste vengono utilizzate, Oracle Workspace Manager 10g gestisce i metadati in modo trasparente all'utente in una sorta di repository.

Con pochi semplici passi è possibile quindi implementare tutti i data mart con i rispettivi attributi. Resta ancora da vedere in che modo sia possibile eseguire l'analisi sui data mart attraverso report grafici. Per questa attività è stato adottato ORACLE Business Intelligence Discoverer with OLAP, uno strumento molto efficiente per realizzare applicazioni web di gestione OLAP che possano generare dinamicamente rapporti grafico-tabellari (fig. 10).

Perciò, un utente attraverso il browser potrà effettuare tutte le operazioni canoniche sul datawarehouse (*drill-down*, *slice-and-dice*, *roll-up*) scegliendo di volta in volta le dimensioni coinvolte nell'analisi, l'eventuale intervallo temporale di analisi e la tipologia di grafico che preferisce (a torta, a barre, etc.). Questa soluzione offre molti vantaggi in termini di interattività con l'utente, pur non sacrificando la portabilità per quanto riguarda il lato client.

ROLAP vs MOLAP

Abbiamo visto che, a fronte di un unico modello concettuale OLAP, le metodologie maggiormente usate sono due: ROLAP e MOLAP. Definiamo quindi come Multidimensional OLAP una applicazione OLAP connessa direttamente a un database multidimensionale; diversamente i Relational OLAP saranno i sistemi formati da un motore OLAP multidimensionale che si avvale di un database relazionale per la memorizzazione dei dati. Non esiste un metodo in assoluto migliore per realizzare un datawarehouse, ma ciascuna metodologia presenta svariati punti di forza.

Punti di forza dei MOLAP:

- in un database MOLAP vi è l'identità tra il concetto di multidimensionalità e come questo viene implementato; ciò favorisce la semplicità e in qualche modo anche la velocità nelle risposte;
- molti sistemi MOLAP possiedono una fornitissima libreria di funzioni finanziarie e statistiche o, ancora, incorporano una sofisticata, ma trasparente, gestione della dimensione tempo;
- per applicazioni che fanno uso intensivo di calcoli (in pratica più simili a un foglio elettronico che a un database) la maggior parte delle volte l'approccio MOLAP si rivela la scelta appropriata;
- i sistemi Multidimensional-OLAP, essendo utilizzati sia per calcoli e presentazioni sia per la modellizzazione, possono incorporare regole aziendali e funzioni personalizzate in modo da aderire più da vicino al modello aziendale.

Punti di forza dei ROLAP:

- non appena la quantità di dati da memorizzare comincia a crescere, la scelta diventa quasi obbligata: la tecnologia relazionale ha alle spalle anni di ricerche nella memorizzazione di grosse quantità di dati, i MOLAP in questi casi devono ricorrere a sofisticati algoritmi per la gestione delle matrici sparse, che ne pregiudicano in qualche modo le prestazioni;
- nel caso di cambiamenti nelle dimensioni, come per esempio la sostituzione dei codici prodotto, un MOLAP dovrà ricalcolare le aggregazioni fino ai dati storici iniziali; se queste operazioni avvengono con una certa frequenza, la scelta di un ROLAP si rivela più appropriata;
- i MOLAP si comportano egregiamente quando la dimensionalità dell'oggetto di analisi è ben definita e possibilmente statica, cambiamenti nella struttura dimensionale richiedono una riorganizzazione fisica del database. Alcuni contesti sono molto dinamici dal punto di vista delle dimensioni, perciò in questi casi è appropriato utilizzare la visione dimensionale dei ROLAP perché questa normalmente viene costruita a ogni analisi.

[6] Informazioni relative ai dati stessi. La definizione di metadati varia in funzione del contesto in cui si applica:

1. nella gestione di un DBMS, rappresentano gli oggetti del database (tabelle, viste, utenti, etc.);
2. nell'acquisizione tramite strumenti ETL (Extract, Transform, Load), rappresentano le modalità di trasformazione dei dati dai sistemi operazionali al datawarehouse;
3. nell'accesso tramite prodotti OLAP, rappresentano il mapping dello schema fisico del database rispetto alla vista ottenibile tramite le query.

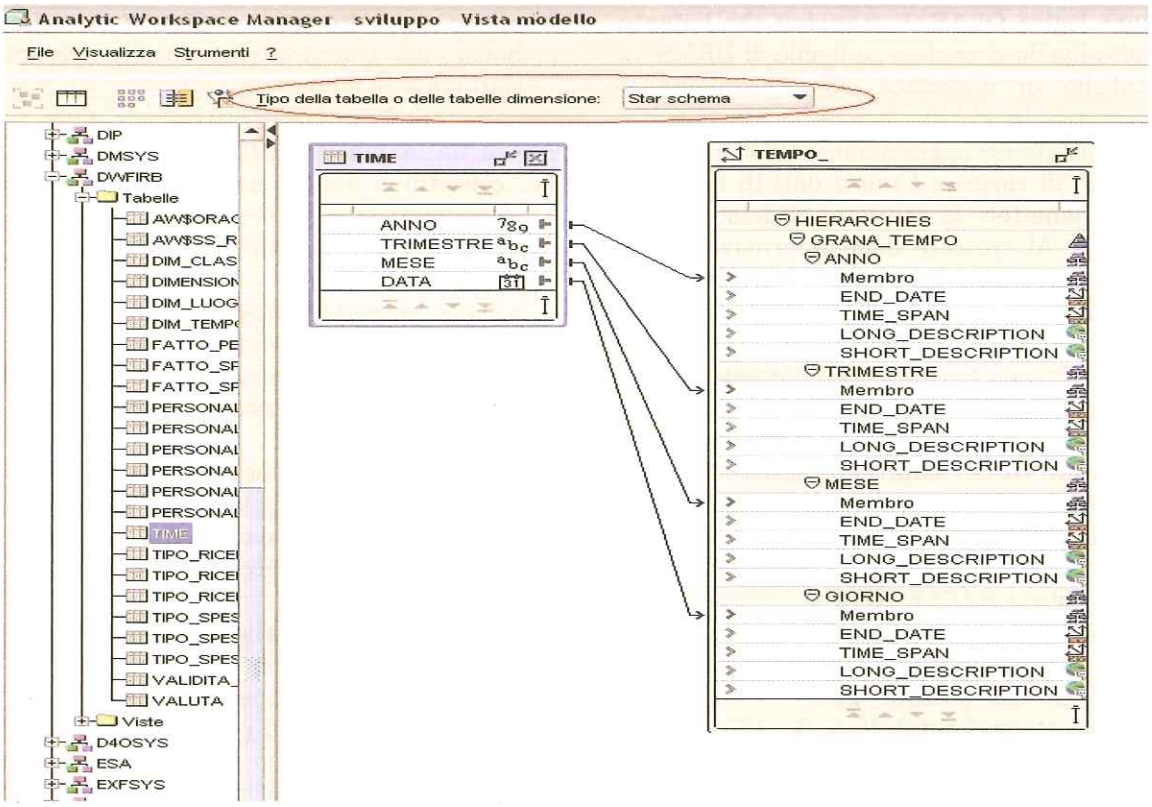

Fig. 9 – Snapshot del mapping di una dimensione in Oracle Workspace Manager 10g

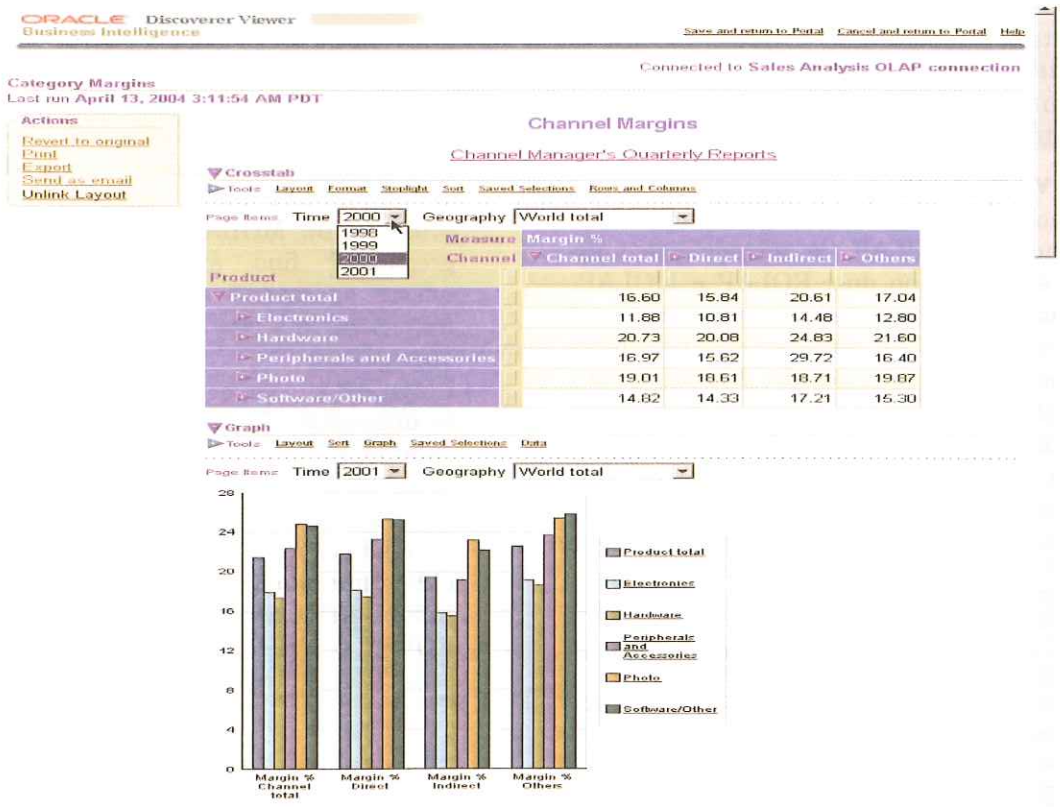

Fig. 10 – Snapshot di una schermata di Oracle Business Intelligence

Bibliografia

[1] B. Potì, "Un progetto FIRB per la valutazione dell'impatto dei finanziamenti pubblici per la ricerca e l'innovazione", *Bollettino del CILEA,* n. 108, ottobre 2007.

[2] W. H. Inmon *Building the Data Warehouse.* Wiley, 2005.

[3] S.Rizzi, M.Golfarelli *Data Warehouse. Teoria e pratica della progettazione.* McGraw Hill, 2004.

[4] P. Fraternali, S. Paraboschi, R. Torlone, P. Atzeni, S. Ceri *Basi di Dati, architetture e linee di evoluzione.* McGraw-Hill, 2003.

[5] M.Ross, R.Kimball *The Data Warehouse Toolkit, second edition* (Wiley). Hoepli, 2003.

[6] J.Caserta, R.Kimball *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Wiley, 2006.

[7] C. Koncilia, R. Wrembel *Data Warehouses and OLAP: Concepts, Architectures and Solutions.* British Library, 2006.

[8] R. Johnson, J. Vlisside, E. Gamma, R. Helm *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley, 2000.

[9] S.Lawande, P.Smith, L.Hobbs, S.Hillson *Oracle Database 10g Data Warehousing.* Elsevier, 2005.